Visible to infrared transfer learning as a paradigm for accessible real-time object detection and classification in infrared imagery

A. Gaus, Yona Falinie, Bhowmik, Neelanjan, Isaac-Medina, Brian K. S., Breckon, Toby

Yona Falinie A. Gaus, Neelanjan Bhowmik, Brian K. S. Isaac-Medina, Toby P. Breckon, "Visible to infrared transfer learning as a paradigm for accessible real-time object detection and classification in infrared imagery," Proc. SPIE 11542, Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies IV, 1154205 (20 September 2020); doi: 10.1117/12.2573968

SPIE.

Event: SPIE Security + Defence, 2020, Online Only

Visible to Infrared Transfer Learning as a Paradigm for Accessible Real-time Object Detection and Classification in Infrared Imagery

Yona Falinie A.Gaus[a], Neelanjan Bhowmik[a], Brian K.S. Isaac-Medina[a], and Toby P. Breckon[a,b]

Department of {Computer Science[a] | Engineering[b]}, Durham University, Durham, UK

ABSTRACT

Object detection from infrared-band (thermal) imagery has been a challenging problem for many years. With the advent of deep Convolutional Neural Networks (CNN), the automated detection and classification of objects of interest within the scene has become popularised due to the notable increases in performance over earlier approaches in the field. These advances in CNN approaches are underpinned by the availability of large-scale, annotated image datasets that are typically available for visible-band (RGB) imagery. By contrast, there is a lack of prior work that specifically targets object detection in infrared-band images, owing to limited datasets availability that stems from more the limited availability and access to infrared-band imagery and associated hardware in general. A viable solution to this problem is transfer learning which can enable the use of such CNN techniques within infrared-band (thermal) imagery, by leveraging prior training on visible-band (RGB) image datasets, and then subsequently only requiring a secondary, smaller volume of infrared-band (thermal) imagery for CNN model fine-tuning. This is performed by adopting an existing pre-trained CNN, pre-optimized for generalized object recognition in visible-band (RGB) imagery, and subsequently fine-tuning the resultant model weights towards our specific infrared-band (thermal) imagery domain task. We use of two state-of-art object detectors, Single Shot Detector (SSD) with a VGG-16 CNN backbone pre-trained on the ImageNet dataset, and You-Only-Look-Once (YOLOV3) with a DarkNet-53 CNN backbone pretrained on the MS-COCO dataset to illustrate our visible-band to infrared band transfer learning paradigm. Exemplar results reported over the FLIR Thermal and MultispectralFIR benchmark datasets show that significant improvements in mAP detection performance to $\{0.804_{MsFIR}, 0.710_{FLIR}\}$ for SSD and $\{0.520_{MsFIR}, 0.308_{FLIR}\}$ for YOLOV3 via the use of transfer learning from initial visible-band based CNN training.

Keywords: object detection, infrared-band imagery, visible-band imagery, CNN, transfer learning, deep learning

1. INTRODUCTION

The use of both 2D visible-band and infrared-band camera imagery within many visual surveillance tasks are well established with numerous solutions spanning target detection, visual tracking and behaviour analytics.[1–4] More broadly, the complementary nature of using both colour and (far, long-wave) infrared has seen them extensively utilised in a range of computer vision applications for several decades.[5] Prior work on 2D visible-band and infrared-band imagery has spanned object tracking,[6,7] pedestrian detection,[8] stereo vision,[9] and autonomous platform deployment.[10] The advantage of thermal infrared-band is that it is not influenced by visible spectrum illumination variations and shadows, and objects-of-interest can be readily distinguished from the scene background which is typically colder. In addition, infrared-band sensing can be used in total darkness, as opposed to the scene illumination requirements of visible-band imagery. Unlike visible-band sensing which operates in a visible light spectrum ranging from {400 - 700 nm}, infrared images are formed by using thermal radiation emitted by all matters with temperatures above absolute zero in the wavelength range from $\{0.1\mu m$ - $100\mu m\}$.[11] In simple terms, almost anything that creates heat can be visible and captured by infrared sensing (Figure 1, right). Infrared sensing is also becoming increasingly low-cost and compact in terms of sensor size, which makes

Further author information: Send correspondence to Yona Falinie A. Gaus.
E-mail: yona.f.binti-abd-gaus@durham.ac.uk

Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies IV, edited by Henri Bouma, Radhakrishna Prabhu, Robert James Stokes, Yitzhak Yitzhaky, Proc. of SPIE Vol. 11542, 1154205
 · doi: 10.1117/12.2573968

it in turn increasingly popular in various applications such as autonomous driving and persistent surveillance. As a result detection, as well as the classification of objects-of-interest in infrared-band imagery, is an important issue to be addressed in terms of the increased use of infrared sensing within a broadening range of intelligent imaging applications.

Figure 1. Example of visible-band colour (RGB) image (left) and far infrared-band (thermal) image (right).

The effectiveness of real-time object detection architectures[12–16] is usually dependent upon a large amount of labelled training imagery. Recently, deep learning has revolutionized the field of computer vision significantly advancing the state-of-the-art in many applications.[17] However, within object detection specifically, most of the efforts have focused on detecting objects-of-interest in standard visible-band (colour, RGB) imagery. Object detection performance in this domain has been significantly improved by using region-based method such as Region-based Convolutional Neural Networks (R-CNN)[18] and Fast Region-based Convolutional Network method (Fast R-CNN)[12] that uses selective search, while its successor, Faster Region-based Convolutional Network method (Faster R-CNN) uses region proposal networks to identify objects-of-interest. An alternative CNN architecture for object detection, You-Only-Look-Once (YOLO)[15] revisits the object detection problem by turning it into regression problem, where the coordinates of the bounding boxes and the class probability for each of those boxes are generated simultaneously. This capability makes YOLO[15] extremely fast in terms of processing time albeit with lesser detection performance than R-CNN counterparts.[19] These aforementioned object detection based CNN methods rely heavily on architectures that have been trained on large-scale visible-band (RGB) imagery datasets such as ImageNet,[20] PASCAL Visual Object Classes (PASCAL-VOC),[21] and Microsoft Common Objects in Context (MS-COCO).[22] Introducing deep learning to object detection within infrared- band (thermal) imagery is significantly hindered by by the absence of such annotated datasets of the same scale and variety. Comparatively the available datasets for infrared-band (thermal) imagery[23,24] are relatively small. In infrared-band (thermal) imagery the lack of such datasets, which is attributable to the lesser prevalence of this sensing modality in general, artificially restricts an equivalent level of CNN success for this spectral band.

Transfer learning is a solution to this problem as it requires only a small volume of infrared-band (thermal) imagery. The theory behind transfer learning is that low-level features learned through training, occurring within the earlier convolutional layers of a CNN model, are common for all objects of interest. It is evident that from the inspection of hidden layers within typical CNN models that each has distinct feature representation related characteristics such that the earlier (CNN) layers can be said to provide general feature extraction capabilities while later layers in the CNN carry information that is increasingly more specific to the original classification task. Often these low-level features are gradient (i.e. edge/texture) based meaning that they occur similarly in varying spectral bands, and hence can be considered to be readily transferable to tasks involving a differing spectral band of imagery to that upon which they were originally trained. By re-using this generalized feature extraction and representation of the earlier layers in a CNN, we can subsequently fined-tune the later layers towards a secondary task within a differing spectral band of imagery. Using this paradigm, we adopt an existing pre-trained CNN, pre-optimized for generalized object recognition in visible-band imagery, and fine-tune the model weights towards our specific infrared-band imagery domain task. In order to emphasize the importance of

transfer learning from one spectral band to another within this paradigm, we compare our use of transfer learning against the use of random weight initialization (i.e. "training from scratch") across a number of contemporary object detection and classification CNN architectures.

2. RELATED WORKS

Object detection and classification in thermal imagery has been an active area of research for a number of years.[25–29] A range of trial work in the literature addresses the task of detecting people and objects in thermal imagery.[28,29] An early method such as a probabilistic template-based approach[28] is proposed for pedestrian detection by using far-infrared (thermal) images. The algorithm is based on the use of a match against probabilistic human shape templates, where it is assumed that human body temperature is warmer compared to background under thermal imagery. Another template-based approach is used to detect people in widely varying thermal imagery.[29] First, a fast screening procedure is undergone using the generalised template to potentially locate a person position. Subsequently, AdaBoost ensemble classifier is used to test the hypothesised person location. Kai *et al.*[30] examined the local feature-based pedestrian detector on thermal data. It uses a combination of multiple cues to find interest points in the images and then uses Speeded-Up Robust Features (SURF) as features to describe these points. These methods show that particular points can be used to recognize specific object parts, i.e., body parts of detected people. In another approach, Loveday *et al.*[31] take another step on pedestrian detection by developing an orthogonal dual camera imaging system to capture parallax-free and well-aligned multispectral images. It is shown that visible and infrared data fusion achieves improved overall performance of foreground object detection than using single-channel visible or infrared information.

With the increasing popularity of deep CNN, several methods have been proposed for applying deep learning methods to thermal imagery.[25–27] A number of studies considered deep learning object detection in infrared-band (thermal) imagery, and in most cases, this research was carried out in the field of autonomous driving, where detecting pedestrian and vehicles are utmost important. Abbot *et al.*[32] are among the first to use a transfer learning approach with the YOLO[33] framework to train a network on high-resolution thermal imagery for classification of pedestrians and vehicles in low-resolution thermal images. The work of Kuramochi *et al.*[34] uses a variant of YOLO, namely Sparse YOLOV2 in FPGA-based pedestrian detector system, due to its efficiency. In the work of Ghose *et al.*[35] trained augmented thermal imagery with saliency maps using Faster R-CNN[13] for the pedestrian detector, reducing the miss rate detection by 13.4% and 19.4% over the baseline in the day and night images respectively. Tumas *et al.*[36] uses F-RCNN[12] and YOLO[33] for pedestrian detectors on far-infrared (FIR) thermal imagery. In another approach Zhang *et al.*[37] uses a Faster-RCNN[13] detector trained on thermal imagery with a super-resolution method to deal with the issue of a small number of pixels that targets at the long-range have. Devaguptapu *et al.*[38] address the data scarcity in thermal imagery by utilizing image-to image translation frameworks[39] to generate pseudo-RGB equivalents of a given thermal imagery, then employing Faster-RCNN[13] for thermal imagery. The work of Chao *et al.*[40] propose one-stage detector namely ThermalDet detector which utilizes all the features in different levels of the feature pyramid extracted by the backbone network, resulting in higher detection accuracy than Faster R-CNN.[40]

Apart from the use of CNN for object detection on thermal imagery, some authors propose the use of CNN for object tracking in thermal imagery. The work by Liu *et al.*[41] transfer pre-trained VGG-Net,[42] trained on visible imagery datasets, to the thermal imagery then uses correlation filter based ensemble tracker for tracking task. Zhang *et al.*[43] address the data scarcity in thermal imagery by utilizing an end-to-end network to generate synthetic thermal imagery from RGB, by using image-to-image translation methods.[39,44] Some authors propose the use of thermal imagery as a complementary of visible imagery by fusing both domains in a deep neural network.[45–47] In the work of Sun *et al.*[45] uses visible and thermal imagery fusion approach, which proposes a fusion-based network for the semantic segmentation of urban scenes, by employing encoder-decoder architecture. Namely RGB-Thermal Fusion Network (RTFNet),[45] the results demonstrate that the superiority of the network, even in challenging lighting conditions. Similar to RTFNet, Multi-spectral Fusion Networks (MFNet) architecture (MFNet) proposed by Ha *et al.*[46] fuses visible and thermal imagery resulting similar or higher accuracy than state-of-the-art segmentation methods such as SegNet.[47]

The effectiveness of deep CNN architectures[12–16] are usually dependent upon a large amount of annotated training imagery. As previously identified, the availability of such resources within the infrared sensing domain

is considerably more limited than that of conventional visible-band imagery and an approach to overcome this is the use of transfer learning. According to Pan *et al.*[48] the aims of transfer learning is *extraction of knowledge from one or more source tasks followed by the application of this knowledge to a target task.* Therefore by using transfer learning, it is possible to benefit from an existing CNN model that has been pre-trained on data resources from a different, yet similar data domain (i.e. visible-band imagery) and then fine tune that model over a more limited quantity of available training data resources in the final target domain (i.e. infrared-band imagery). To emphasize the importance of transfer learning from one spectral band to another within this paradigm, we compare our use of transfer learning against the use of random weight initialization (i.e. "training from scratch") across a number of contemporary object detection and classification CNN architectures.

3. METHODOLOGY

The motivation of this work is to address the problem of data scarcity of annotated infrared-band (thermal) imagery. The key idea is to use transfer learning which can enable the use of such CNN techniques within infrared-band imagery, by leveraging prior training on visible-band image datasets, and then subsequently only require a secondary, smaller volume of infrared-band imagery for CNN model fine-tuning. In the following section, we detail the various elements of experimental methodology to show the effectiveness of transfer learning within context of object detection and classification within infrared-band imagery.

3.1 Datasets

To the best of our knowledge, the FLIR[24] and Multispectral[23] datasets are the only thermal imagery datasets of notable size that also contain a wide variety of object-of-interest annotations. There are several datasets such as KAIST[49] and BU-TIV[50] datasets but they depict only one object class (pedestrian).

Multispectral[23]: The UTokyo dataset contains a total of 7,512 images (3,740: during day and 3,772: during night), which are taken in a university environment at 1 fps using visible-band (RGB colour), Far Infrared (FIR), Mid Infrared (MIR), and Near Infrared (NIR) cameras (as specified within[23]). In this work, we utilise only the Far Infrared images (FIR), taken by Nippon Avionics, InfReC R500, as a dataset for object detection (denoted as MultispectralFIR). Five classes, {*pedestrians, car, bicycle, colour-cone, car-stop*}, are labelled in this dataset consisting of 6,008 training images and 1,502 test images. From Table 1, the class imbalance issue is extremely severe in the MultispectralFIR dataset, challenging the performance of proposed approaches. There is no information provided in terms of occlusion between object, however the size of object-of-interest, along with the average object instance size are presented in Table 1.

Table 1. MultispectralFIR infrared-band dataset statistics.

Statistics	Multispectral FIR				
	Pedestrian	Car	Bike	Colour-cone	Car-stop
Avg. (h×w) px	85×34	105×113	99×112	75×41	73×55
Max (h×w) px	475×487	470×627	391×637	180×287	389×640
Min (h×w) px	2×10	1×2	9×5	1×4	9×2
Training instances	10,452	4,070	2,544	2,245	4,306
Test instances	2,769	972	621	634	1,076
Number of Images	Training			Test	
	5,557			1385	

FLIR[24]: The FLIR thermal dataset provides annotated thermal images and non-annotated RGB images for training and testing of object detection and classification approaches. The dataset is acquired via a visible and thermal camera mounted on a vehicle. All videos are taken on the streets and highways in Santa Barbara, California, USA from November to May. Videos are taken under generally clear-sky conditions during day and night. Thermal images are acquired with a FLIR Tau2 (13 mm f/1.0, 45-degree HFOV and 37-degree VFOV). RGB images are acquired with a FLIR BlackFly at 1280×512 (4-8 mm f/1.4-16 megapixel lens with the FOV set to match Tua2). Both cameras are operated in their default configuration as supplied by the manufacturer. The cameras are in a single enclosure 1.9 +/-0.1 inches apart from each other. Images are captured via USB3

video using FLIR-proprietary software. The majority of the 10,228 thermal images are sampled at a rate of two images per second (native videos are 30 frames per second of video). A minority of images that are acquired in less object-rich environments, are sampled at a rate of one image per second. In this study, we only concentrate on the imagery acquired by thermal camera where we use the 10,228 images provided for this task. The details of training and testing image sets follow exactly FLIR[24] protocol, where validation images are not explicitly defined in the data set definition, but we randomly select 20% subset of the training examples to form the validation set. The detail of the dataset statistics are presented in Table 2.

Table 2. FLIR infrared-band dataset statistics.

Statistics	FLIR			
	Person	Bicycle	Car	Dogs
Avg. (h×w) px	40×15	41×31	40×46	26×27
Max. (h×w) px	429×288	141×296	318×493	78×104
Min (h×w) px	5×3	10×3	4×9	6×11
Training instances	35,520	4,202	44,126	192
Test instances	5,779	471	5,432	14
Number of Images	Training		Test	
	9,620		1,366	

3.2 CNN Architectures

We use YOLOV3[16] and SSD[51] as the primary CNN architectures for object detection based on performance within prior comparative work in the field.[52–54] Each of the down-selected CNN architectures (YOLOV3,[16] and SSD[51]) are depicted in Figures 2 & 3. Transfer learning is employed in each CNN architecture, by leveraging a trained sets of network weights for related object detection and classification task or domain. In this task, we leverage the CNN parametrisation of an existing fully trained network on a generic object class problem,[20] as a starting point for optimisation towards identifying objects-of-interest in thermal imagery.

YOLOV3:[16] the You Only Look Once (YOLOV3) architecture operates by introducing $S \times S$ grid to image feature space and bounding boxes are proposed for each grid cell. Motivated by the Faster R-CNN[13] architecture, it uses a set of anchor boxes in each grid cell to parameterise the bounding box dimensions. Instead of using hand-crafted anchor dimensions, it employes k-means to cluster the dimensions on the training set bounding boxes and use them as priors for the anchor dimensions. By using DarkNet-53,[55] the output units are generated by an additional convolutional layer. Darknet-53[55] is a bespoke feature extraction network proposed by the YOLOV3[16] architecture that extracts the features at three different scales, similar to feature pyramid networks (FPN),[56] as shown in Figure 2. In order to improve detection for small objects, it adds pass through layers by bringing mid-level features in order to gain fine grained insight. Finally, candidate object detections are passed to a stage of to Non Maximal Suppression (NMS), which performs detection filtering based on the Intersection over Union (IoU) among detection boxes. NMS selects the box which has the highest confidence score as the detection result, and then it discards other candidate boxes whose Intersection-over-Union (IoU) value with the selected box greater than a given overlap threshold, as shown in Figure 2.

SSD:[51] the Single Shot MultiBox Detector architecture is based on a feed-forward convolutional network, followed by a non-maximum suppression stage to generate the final detection output. SSD[51] consists of two components: a backbone CNN and a SSD head. The overall SSD architecture builds on the VGG-16[42] CNN backbone, but discards the fully connected layers. Instead of using the original VGG-16 fully connected layers, a set of auxiliary convolutional layers (from the 6th convolutional layer, *Conv6*, and subsequently - Figure 3) are added, thus enabling it to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer, as shown in Figure 3. Additionally, SSD combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes, as shown in the red color in Figure 3. The SSD architecture is simple compared to two architectures that require additional region proposal[13] or anchor box[16] formulations because it completely encapsulates all computation in a single network. The SSD head is one or more convolutional layers added to this backbone and the outputs are the sets of bounding boxes and classes of objects in the spatial location of the final layers activations, as shown in final layer of Figure 3.

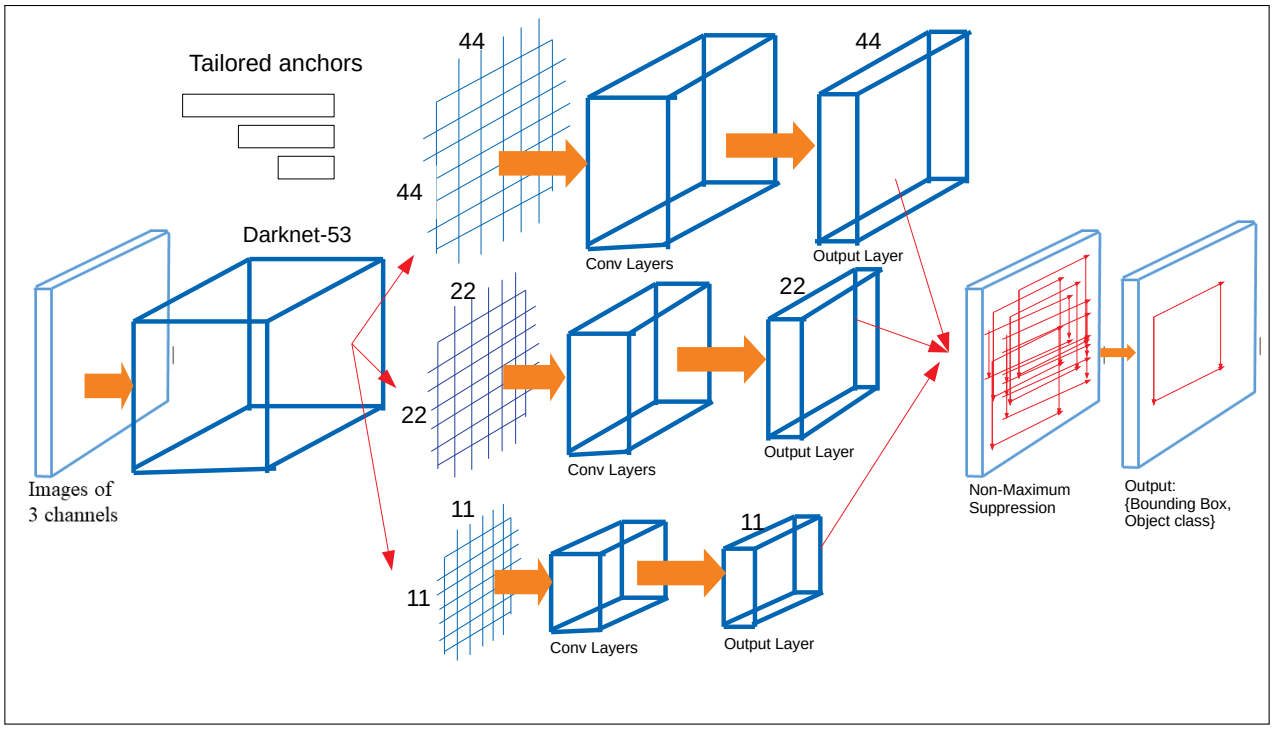

Figure 2. YOLOV3 architecture of the CNN based detection approach evaluated.

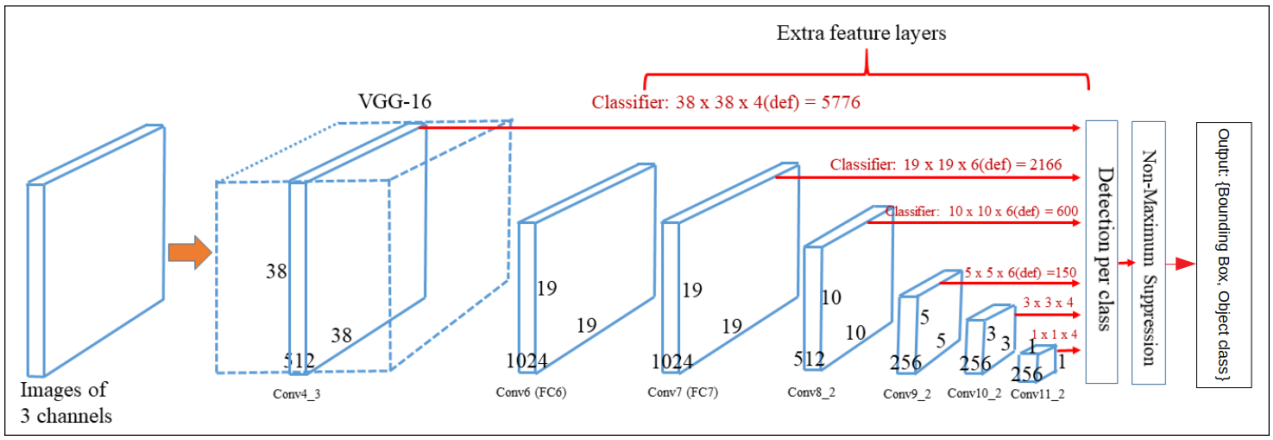

Figure 3. SSD architecture of the CNN based detection approach evaluated.

3.3 Experimental Setup

The object detection architectures (Section 3.3) evaluated in this work are implemented in PyTorch[57] with pre-trained weights from the generic ImageNet[20] object detection and classification task. The YOLOV3[16] is trained using DarkNet-53 backbone[55] with following parameter configuration: backpropagation optimisation performed via by Stochastic Gradient Descent (SGD), learning rate = 0.001, batch size = 8. For SSD,[51] we use VGG-16[42] backbone with following training configuration: backpropagation optimisation performed via by SGD, learning rate = 0.001, batch size = 8.

Our experimental setup comprises of two datasets, FLIR and MusltispectralFIR (Section 3.1), for automatic detection and classification of objects-of-interest from an infrared-band image. Through the transfer learning paradigm, training is initialised with weights from the pre-trained network (Section 3.2) for the visible-band ImageNet classification reference task. We also compare our results against "training from scratch" where all the parameters or weights in the pre-trained network are randomly initialised via normal distribution.

3.4 Performance Evaluation

In order to evaluate the architecture on the task of object detection and classification, we must first determine how well the model predicts the location of the object. Usually, this is done via a 2D bounding box around the object of interest within the image. For all of these cases, this in-image localisation task is typically evaluated on the Intersection over Union (IoU) threshold. Object detections are determined to be true or false depending upon a threshold on the IoU value. In this work, we follow the protocol of MS-COCO challenge[22] whereby the IoU threshold is set throughout the range of 0.5 to 0.95 in steps of 0.05.

At each confidence threshold, we can measure the *Precision* and *Recall* of the detection approach, from which we can plot the *Precision-Recall* curve. The better the performance of a given architecture, the higher the precision and recall at each and every point on the curve. Subsequently, we can summarize the performance of the architecture with one metric, by taking the area under the curve. This gives us a statistic, commonly known as *Average Precision (AP)* between 0 and 1, where higher is better.

In order to calculate AP, we calculate the area of IoU between the given ground truth and detected bounding box for each target object (i.e. object of interest) as:

$$\Psi(B_{gt_i}, B_{dt_i}) = \frac{Area(B_{gt_i} \cap B_{dt_i})}{Area(B_{gt_i} \cup B_{dt_i})} \tag{1}$$

where B_{gt_i} and B_{dt_i} are ground truth and detected bounding box for detection i, respectively. Assuming each detection, i, is unique, and denoting the area as $\Psi(B_{gt_i}, B_{dt_i})$, we then threshold it by the range of $\theta = .50 : .05 : .95$ giving the logical b_i, where:

$$b_i = \begin{cases} 1, & 0.50 < \Psi(B_{gt_i}, B_{dt_i}) < 0.95 \\ 0, & a_i < 0.50 \end{cases} \tag{2}$$

Given both true positive and false positive as TP_i and FP_i, where:

$$\begin{aligned} TP_i =& TP_{i-1} + b_i \\ FP_i =& FP_{i-1} + (1 - b_i) \end{aligned} \tag{3}$$

The *Precision* P_i and *Recall* R_i curves can be calculated as:

$$\begin{aligned} P_i =& \frac{TP_i}{TP_i + FP_i} \\ R_i =& \frac{TP_i}{n_p} \end{aligned} \tag{4}$$

where n_p is the number of positive samples. We can calculate *Average Precision (AP)* based on the area under the curve of *Precision* versus *Recall*:

$$AP = \sum_{i}^{n_d} P_i \triangle R \tag{5}$$

where $\triangle R$ is the *Recall* levels in an increasing order where the *Precision* is first interpolated. Subsequently, we can get the value of mAP by averaging AP values for all classes, C:

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c \tag{6}$$

where C represents the number of target object types (classes) that the detection approach has been trained to detect.

4. RESULTS

The statistical results presented in Tables 3 - 6 consist of Average Precision (AP) for each class and mean Average Precision (mAP) across all classes, as detailed in Section 3.4. The performance statistics for each category of object is divided by the two architectures considered: YOLOV3[16] and SSD[51] (as described in Section 3.2). The bold value highlighted within Tables 3 - 6 indicates maximal performance achieved for each CNN architecture, respectively. Following established convention in the machine learning an object detection literature, the results are presented normalised to unit range, 0 - 1, with a maximal (best performing) result indicated by a higher value (for AP/mAP).

Table 3. MultispectralFIR Dataset: Training via transfer learning.

Model	Backbone	Average Precision (AP)					mAP
		Pedestrian	Car	Bike	Colour-cone	Car-stop	
SSD	VGG-16	0.824	0.896	0.805	0.741	0.753	**0.804**
YOLOV3	Darknet-53	0.584	0.802	0.575	0.505	0.477	0.589

Table 4. MultispectralFIR Dataset: Training via random initialisation.

Model	Backbone	Average Precision (AP)					mAP
		Pedestrian	Car	Bike	Colour-cone	Car-stop	
SSD	VGG-16	0.822	0.873	0.750	0.701	0.688	**0.767**
YOLOV3	Darknet-53	0.465	0.155	0.033	0.080	0.023	0.151

Table 3 presents resulting evaluations for the MultispectralFIR dataset, trained with the transfer learning paradigm. In terms of transfer learning upon individual CNN architectures, it can be observed that SSD[51] architecture with VGG-16[42] pre-trained backbone provides maximum object detection accuracy (i.e high AP in all class). It can be seen that object classes that have more examples within the training dataset, such as *pedestrian*, *car*, and *bicycle*, have superior AP in both approaches. However, object classes *colour-cone* and *colour-stop* have relatively lower AP despite having sufficient number of instances in training sets (Table 1). This may contribute to the fact that SSD performs better (mAP: 0.804) for large size object classes, by employing multiscale features and efficient throughput in terms of frame rate. In comparison to SSD, the YOLOV3[16] architecture with DarkNet-53[55] backbone provides a somewhat average performance overall (mAP: 0.589), with the exception of AP for *car* owing to the large number of examples of this object within the training dataset (Table 1).

We compare our use of transfer learning (Table 3) against the use of random weight initialisation (i.e. "training from scratch") across two contemporary object detection and classification CNN architectures (Table 4). Overall, it indicates that for each CNN architecture, transfer learning by fine-tuning using pretrained weights as an initialisation is better than random weights initialisation even if the target task (infrared-band imagery) is very different from the source task (visual-band imagery). However it is worth noting that training via random initialisation provides somewhat comparable performance to training with transfer learning (Table 3 & 4, SSD). The results for object classes that have a large number of training instances in the dataset such as *pedestrian* and *car* manage to achieve high detection accuracy.

The statistical evaluation presented in Table 3 and 4 translates to the confusion matrices presented within Figure 4 (a/b) and Figure 4 (c/d). As for training from transfer learning (Table 3 - Figure 4 (a/b)), strong True Positive (TP) performance occurs within SSD (a) and YOLOV3 (b) confusion matrix for object class *pedestrian* and *car* with these object class labels again being dominant within the training dataset. We can see quite a number of False Positive (FP) occurrence (off-diagonal) across object class of *colour cone* and *car stop* being misclassified as background. Overall, strong TP performance (diagonal) and low FP occurrence (off-diagonal) on both confusion matrix (a/b) in Figure 4 reflects the strong AP and mAP results from Table 1. As for training from random initialisation (Table 4, Figure 4 (c/d)), it is observed that strong TP performance occurs within SSD (c) confusion matrix for object class *pedestrian*, *car* and *bike*, again being dominant within the training datasets. However, strong FP occurrence (off-diagonal) on YOLOV3 (d) where we can observe that there is significant portion of *bike* misclassified as *car-stop*, *car-stop* misclassified as background, and *color-cone*

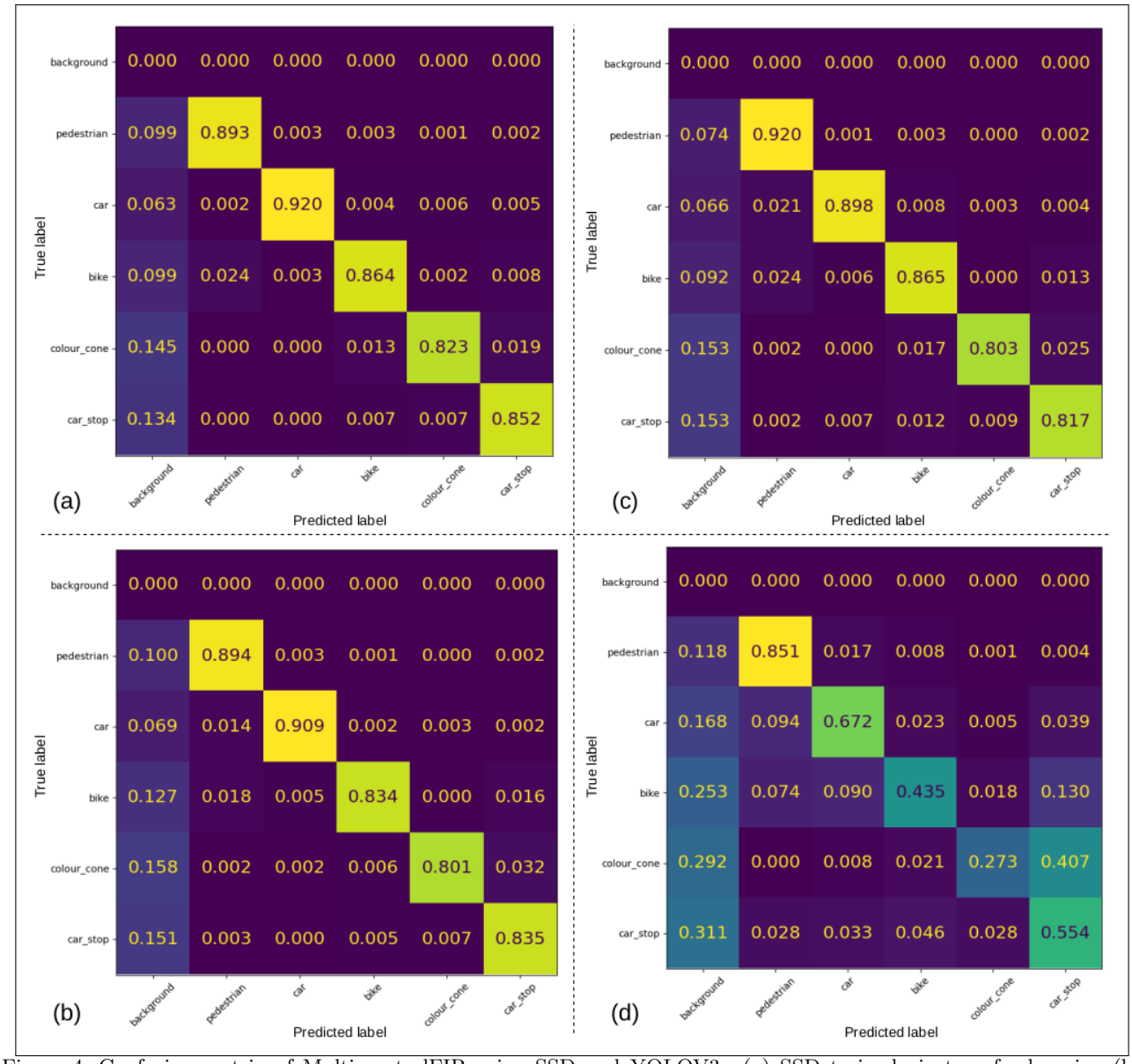

Figure 4. Confusion matrix of MultispectralFIR using SSD and YOLOV3. (a) SSD trained via transfer learning (b) YOLOV3 trained via transfer learning (c) SSD trained via random initialisation (d) YOLOV3 trained via random initialisation.

misclassified as *car-stop*. This is probably attributable to the fact that most of the *bike* objects occur near to the *car-stop*, and *colour-cone* is very often adjacent to *car-stop* which complicates detection. Overall, this indicates that training with transfer learning from pre-trained networks can achieve better performance than training from random initialisation for this cross spectral-domain task. Qualitative results for the MultispectralFIR dataset are illustrated over a number of examples in Figure 6.

A consistent result is observed in Tables 5 and 6 for the FLIR dataset, trained with transfer learning architecture as well as random initialisation. Overall, both architectures in Table 5 (transfer learning) manage to outperform training with random initialisation in Table 6. In comparison between the SSD and YOLOV3 architectures in terms of class-wise, the SSD architecture achieves superior results on object class *person* and *car* for SSD and object class *car* for YOLOV3, owing to classes that have a large number of training instances in the dataset (Table 2). Whilst reasonable performance is achieved for most of the object classes by YOLOV3

Table 5. FLIR Dataset: Training via transfer learning.

Model	Backbone	Average Precision (AP)				mAP
		Person	Bicycle	Car	Dogs	
SSD	VGG-16	0.619	0.461	0.851	0.149	**0.520**
YOLOV3	Darknet-53	0.332	0.345	0.554	0.000	0.308

Table 6. FLIR Dataset: Training via random initialisation.

Model	Backbone	Average Precision (AP)				mAP
		Person	Bicycle	Car	Dogs	
SSD	VGG-16	0.599	0.460	0.843	0.103	**0.501**
YOLOV3	Darknet-53	0.237	0.220	0.571	0.000	0.257

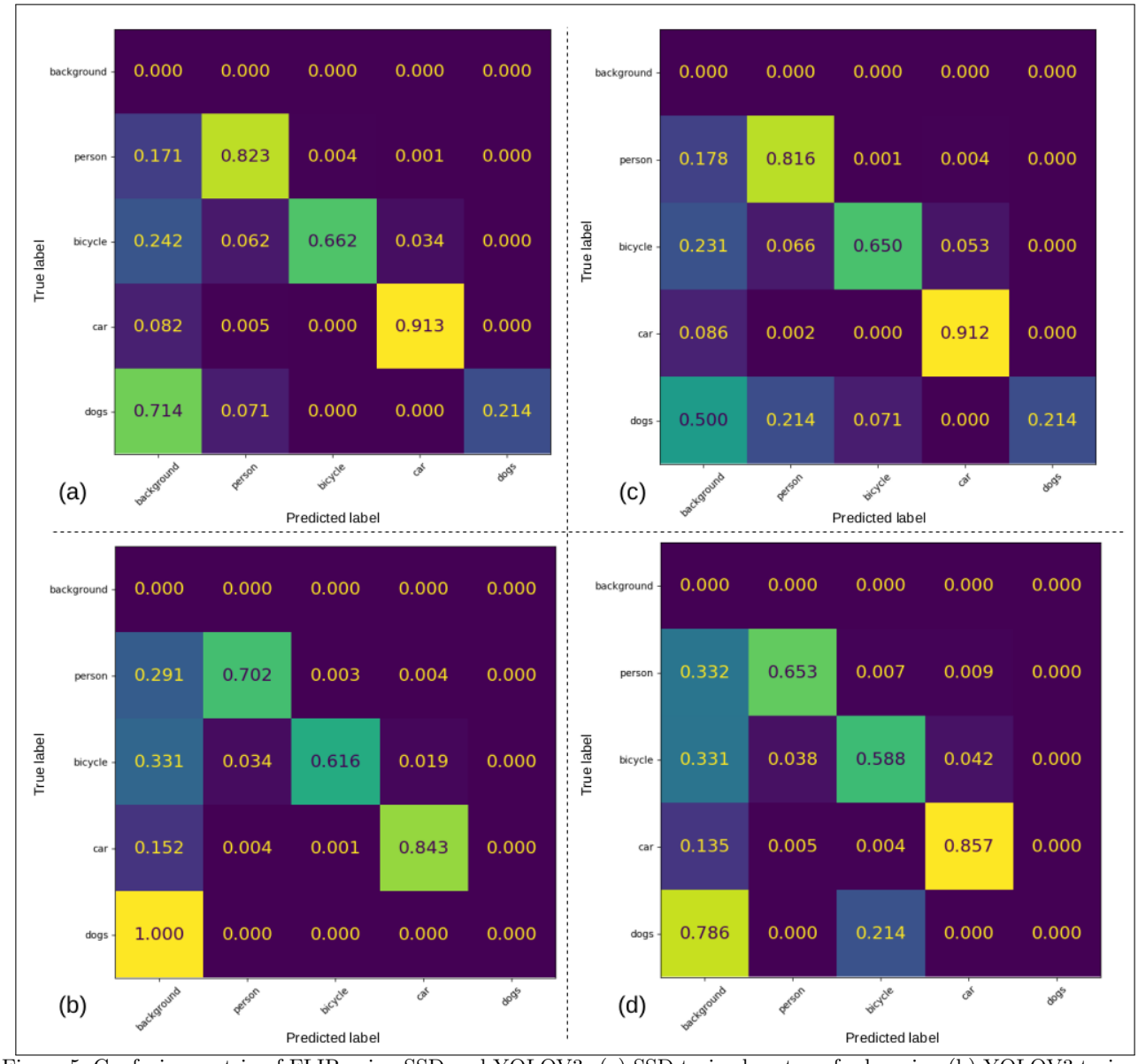

Figure 5. Confusion matrix of FLIR using SSD and YOLOV3. (a) SSD trained on transfer learning (b) YOLOV3 trained on transfer learning (c) SSD trained on random initialisation (d) YOLOV3 trained on random initialisation.

Figure 6. Exemplar detection of MultispectralFIR dataset using SSD and YOLOV3.

architecture, the resultant model failed to be able to detect any of the examples of object class *dogs*.

By contrast, SSD manages to detect object types that have a low number of labelled examples within the training set, such as *dogs*, in both training with transfer learning and random initialisation (Tables 5 & 6, SSD). The detection of small objects demonstrates the superiority of SSD architecture for small object detection. It is observed that even in random initialisation, the only class label that has large instances in the training sets such as *person* and *car* manage to achieve such comparable performance as in transfer learning. It is probably attributable to the imbalance within the dataset where a relatively small number of examples for the class *bicycle* and *dogs* which are present. Moreover, as the calculation of the mAP is always carried out as an average across all of the dataset object class types, the impact of the zero AP for the single object class *dogs* directly contributes to YOLOV3 having the lowest mAP (mAP:0.308, mAP:0.257) for performance within this evaluation (Table 5 - 6, YOLOV3). Such results are directly indicative of the impact of dataset imbalance upon the training process.

The statistical evaluation presented in Tables 5 and 6 translates to the confusion matrices presented within Figure 5 (a/b) and Figure 5 (c/d). As for training from transfer learning (Table 5 - Figure 5 (a/b)), strong TP performance occurs within SSD (a) and YOLOV3 (b) confusion matrix for object class *person* and *car* with these object class labels again being dominant within the training dataset. We can see quite a number of FP occurrence (off-diagonal) across object class of *person*, *bicycle* and *dogs* being misclassified as background. As for training from scratch (Table 6, Figure 5 (c/d)), it is observed that strong TP performance occurs within SSD (c) confusion matrix for object class *person* and *car* again being dominant within the training datasets. However, strong FP occurrence (off-diagonal) on YOLOV3 (d) where we can observe that there is significant portion of

bicycle and *dogs* misclassified as background. This is probably attributable to the imbalance within the dataset and a relatively small number of examples for the class *bicycle* and *dogs* which are present. Finally, qualitative results for the FLIR dataset are illustrated over a number of examples in Figure 7.

Figure 7. Exemplar detection in FLIR dataset using SSD and YOLOV3.

5. CONCLUSION

This work explores the use of transfer learning for CNN based object detection techniques within infrared-band (thermal) imagery, by leveraging prior training on visible-band (RGB) image datasets, and then subsequently only requiring a secondary, smaller volume of infrared-band (thermal) imagery for CNN model fine-tuning. We use existing pre-trained models, which are pre-optimised for generalised object detection in visible-band imagery, and subsequently apply transfer learning for object detection in infrared-band imagery. The proposed transfer learning approach achieves superior performance on SSD (mAP: 0.804) as well as YOLOV3 (mAP: 0.589) respectively. This offers a significant improvement over training from random initialisation on SSD (mAP: 0.767) and YOLOV3 (mAP: 0.151) respectively, on the MultispectralFIR dataset. As for FLIR dataset, the proposed transfer learning approach achieves higher on SSD performance (mAP: 0.520) as well as YOLOV3 (mAP: 0.308) over training from random initialisation on SSD (mAP: 0.501) and YOLOV3 (mAP: 0.257) respectively. Whilst we observe that both detection architectures struggle to detect object types with low number of samples within the training dataset, a potential area for future work is the evaluation of the impact of imbalanced datasets within object detection training for these and other object detection and classification architectures in general.

REFERENCES

[1] Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., and Shah, M., "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(7), 1442–1468 (2013).

[2] Kundegorski, M. E. and Breckon, T. P., "A photogrammetric approach for real-time 3d localization and tracking of pedestrians in monocular infrared imagery," in [*Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology XI*], **9253**, 92530I, International Society for Optics and Photonics (2014).

[3] Kundegorski, M. E., Akçay, S., de La Garanderie, G. P., and Breckon, T. P., "Real-time classification of vehicles by type within infrared imagery," in [*Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*], **9995**, 99950T, International Society for Optics and Photonics (2016).

[4] Kundegorski, M. E. and Breckon, T. P., "Posture estimation for improved photogrammetric localization of pedestrians in monocular infrared imagery," in [*Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XI; and Optical Materials and Biomaterials in Security and Defence Systems Technology XII*], **9652**, 96520F, International Society for Optics and Photonics (2015).

[5] Lin, S.-S., "Extending visible band computer vision techniques to infrared band images," (2001).

[6] Colantonio, S., Benvenuti, M., Di Bono, M., Pieri, G., and Salvetti, O., "Object tracking in a stereo and infrared vision system," *Infrared physics & technology* **49**(3), 266–271 (2007).

[7] Torabi, A., Massé, G., and Bilodeau, G.-A., "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding* **116**(2), 210–221 (2012).

[8] Krotosky, S. J. and Trivedi, M. M., "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems* **8**(4), 619–629 (2007).

[9] Pinggera, P., Breckon, T., and Bischof, H., "On cross-spectral stereo matching using dense gradient features," in [*Proc. British Machine Vision Conference*], 526.1–526.12, BMVA (September 2012).

[10] Breckon, T. P., Gaszczak, A., Han, J., Eichner, M. L., and Barnes, S. E., "Multi-modal target detection for autonomous wide area search and surveillance," in [*Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X*], **8899**, 889913, International Society for Optics and Photonics (2013).

[11] Meseguer, J., Pérez-Grande, I., and Sanz-Andrés, A., [*Spacecraft thermal control*], Elsevier (2012).

[12] Girshick, R., "Fast r-cnn," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1440–1448 (2015).

[13] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in [*Advances in Neural Information Processing Systems*], 91–99 (2015).

[14] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of the IEEE Intl. conference on computer vision*], 2980–2988 (2017).

[15] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 779–788 (2016).

[16] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767* (2018).

[17] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in Neural Information Processing Systems*], 1097–1105 (2012).

[18] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 580–587 (2014).

[19] Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S. Z., "Single-shot refinement neural network for object detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 4203–4212 (2018).

[20] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*Conference on Computer Vision and Pattern Recognition*], 248–255, IEEE (2009).

[21] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision* **88**(2), 303–338 (2010).

[22] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context," in [*European Conference on Computer Vision*], 740–755, Springer (2014).

[23] Takumi, K., Watanabe, K., Ha, Q., Tejero-De-Pablos, A., Ushiku, Y., and Harada, T., "Multispectral object detection for autonomous vehicles," in [*Proceedings of the on Thematic Workshops of ACM Multimedia 2017*], 35–43 (2017).

[24] FLIRSystems, "FLIR Thermal Datasets for Algorithm Training." https://www.flir.co.uk/oem/adas/dataset/.

[25] Peng, M., Wang, C., Chen, T., and Liu, G., "NIRFaceNet: A convolutional neural network for near-infrared face identification," *Information* **7**(4), 61 (2016).

[26] Lee, E. J., Ko, B. C., and Nam, J.-Y., "Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night," *Infrared Physics & Technology* **76**, 261–270 (2016).

[27] Rodger, I., Connor, B., and Robertson, N. M., "Classifying objects in lwir imagery via cnns," in [*Electro-Optical and Infrared Systems: Technology and Applications XIII*], **9987**, 99870H, Intl. Society for Optics and Photonics (2016).

[28] Bertozzi, M., Broggi, A., Gomez, C. H., Fedriga, R., Vezzoni, G., and DelRose, M., "Pedestrian detection in far infrared images based on the use of probabilistic templates," in [*Intelligent Vehicles Symposium*], 327–332, IEEE (2007).

[29] Davis, J. W. and Keck, M. A., "A two-stage template approach to person detection in thermal imagery," in [*2005 Seventh IEEE Workshops on Applications of Computer Vision-Volume 1*], **1**, 364–369, IEEE (2005).

[30] Jungling, K. and Arens, M., "Feature based person detection beyond the visible spectrum," in [*2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*], 30–37, IEEE (2009).

[31] Loveday, M. and Breckon, T. P., "On the impact of parallax free colour and infrared image co-registration to fused illumination invariant adaptive background modelling," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 1186–1195 (2018).

[32] Abbott, R., Del Rincon, J., Connor, B., and Robertson, N., "Deep object classification in low resolution lwir imagery via transfer learning," in [*Proceedings of 5th IMA Conference on Mathematics in Defence*], **2** (2017).

[33] Redmon, J. and Farhadi, A., "Yolo9000: better, faster, stronger," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 7263–7271 (2017).

[34] Kuramochi, R., Shimoda, M., Sada, Y., Sato, S., and Nakahara, H., "Fpga-based accurate pedestrian detection with thermal camera for surveillance system," in [*2019 International Conference on ReConFigurable Computing and FPGAs (ReConFig)*], 1–5, IEEE (2019).

[35] Ghose, D., Desai, S. M., Bhattacharya, S., Chakraborty, D., Fiterau, M., and Rahman, T., "Pedestrian detection in thermal images using saliency maps," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 0–0 (2019).

[36] Tumas, P., Jonkus, A., and Serackis, A., "Acceleration of hog based pedestrian detection in fir camera video stream," in [*Open Conference of Electrical, Electronic and Information Sciences (eStream)*], 1–4, IEEE (2018).

[37] Zhang, H., Luo, C., Wang, Q., Kitchin, M., Parmley, A., Monge-Alvarez, J., and Casaseca-De-La-Higuera, P., "A novel infrared video surveillance system using deep learning based techniques," *Multimedia Tools and Applications* **77**(20), 26657–26676 (2018).

[38] Devaguptapu, C., Akolekar, N., Sharma, M. M., and Balasubramanian, V. N., "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in [*Conference on Computer Vision and Pattern Recognition Workshops*], 1029–1038, IEEE (2019).

[39] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*Proceedings of the IEEE International Conference on Computer Vision*], 2223–2232 (2017).

[40] Cao, Y., Zhou, T., Zhu, X., and Su, Y., "Every feature counts: An improved one-stage detector in thermal imagery," in [*International Conference on Computer and Communications*], 1965–1969, IEEE (2019).
[41] Liu, Q., Lu, X., He, Z., Zhang, C., and Chen, W.-S., "Deep convolutional neural networks for thermal infrared object tracking," *Knowledge-Based Systems* **134**, 189–198 (2017).
[42] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," in [*International Conference on Learning Representations*], (2015).
[43] Zhang, L., Gonzalez-Garcia, A., van de Weijer, J., Danelljan, M., and Khan, F. S., "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Transactions on Image Processing* **28**(4), 1837–1850 (2018).
[44] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1125–1134 (2017).
[45] Sun, Y., Zuo, W., and Liu, M., "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters* **4**(3), 2576–2583 (2019).
[46] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., and Harada, T., "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in [*International Conference on Intelligent Robots and Systems*], 5108–5115, IEEE (2017).
[47] Badrinarayanan, V., Handa, A., and Cipolla, R., "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293* (2015).
[48] Pan, S. J. and Yang, Q., "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering* **22**(10), 1345–1359 (2009).
[49] Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I., "Multispectral pedestrian detection: Benchmark dataset and baseline," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1037–1045 (2015).
[50] Wu, Z., Fuller, N., Theriault, D., and Betke, M., "A thermal infrared video benchmark for visual analysis," in [*IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 201–208 (2014).
[51] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "Ssd: Single shot multibox detector," in [*European Conference on Computer Vision*], 21–37, Springer (2016).
[52] Akcay, S., Kundegorski, M. E., Willcocks, C. G., and Breckon, T. P., "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery," *IEEE Transactions on Information Forensics and Security* **13**(9), 2203–2215 (2018).
[53] Gaus, Y., Bhowmik, N., Akcay, S., and Breckon, T., "Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within x-ray security imagery," in [*International Conference on Machine Learning Applications*], IEEE (December 2019).
[54] Bhowmik, N., Q., W., Gaus, Y., Szarek, M., and Breckon, T., "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composite x-ray imagery," in [*British Machine Vision Conference Workshops*], 1–8, BMVA (September 2019).
[55] Redmon, J., "Darknet: Open source neural networks in c." http://pjreddie.com/darknet/ (2013–2016).
[56] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature pyramid networks for object detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 2117–2125 (2017).
[57] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems*], 8024–8035 (2019).